KM 30(1989)
pag 57 - 65

INCREASING PRECISION OR REDUCING EXPENSE IN REGRESSION EXPERIMENTS BY USING INFORMATION FROM A CONCOMITANT VARIABLE

B. Engel*), P. Walstra**)

SUMMARY

A method is presented to increase precision or reduce expense in regression experiments by partly replacing expensive observations on the variable of interest by cheaper observations on a concomitant variable. An application to the prediction of the lean meat percentage of a pig carcass is given.

1. INTRODUCTION

A method to improve precision of estimates or reduce expense without loss of precision relative to direct regression is presented. The method is potentially useful when observations on the dependent variable are more expensive than observations on some related concomitant variable. It consists of collecting the expensive observations on the dependent variable for a subset of the experimental units only, while observing the concomitant variable for all units. Typically the subset will be considerably smaller than the entire sample.

An important application is the prediction of the lean meat percentage of a pig carcass from objective carcass measurements. The actual lean meat content of a carcass may be determined by complete dissection, which is very expensive, or alternatively, by a less accurate but cheaper incomplete dissection method.

A practical example will be discussed where carcass dissections were carried out in the Netherlands, mainly according to the standard method of the Research Institute for Animal Production 'Schoonoord' (incomplete dissection) and partly by the EC-reference method (complete dissection). A prediction formula for the EC-reference lean meat percentage with carcass measurements obtained with the Hennessy Grading Probe as explanatory variables was

Key-words: regression; double-sampling; lean meat percentage; classification of pigs.

---

*) Agricultural Mathematics Group, P.O.Box 100, 6700 AC Wageningen, The Netherlands (tel.: 08370-19100)

**) Research Institute for Animal Production 'Schoonoord', P.O.Box. 501 3700 AM Zeist, The Netherlands

constructed by linear regression. Both the information contained in the complete and incomplete dissections was used to estimate the regression coefficients and residual variance.

2. ESTIMATION

Suppose that for N independent experimental units a concomitant variable Y_* and explanatory variables $x_1, x_2, \ldots, x_k$ are observed. For n ($n \leq N$) out of the N units additionally the variable of interest Y is observed.

For notational convenience and without loss of generality it is mainly assumed that k=1 and $x_1 = x$.
Interest lies in the linear regression of Y on x:

$$E(Y|x) = A + Bx \qquad \ldots(1)$$

$$Var(Y|x) = \sigma^2.$$

It is further assumed that

$$E(Y_*|x) = a + bx, \quad var(Y_*|x) = \sigma_0^2 \qquad \ldots(2)$$

$$E(Y|x,Y_*) = \alpha + \beta x + \gamma Y_*, \quad var(Y|x,Y_*) = \sigma_1^2.$$

Throughout this paper inference will be conditional upon the values of the explanatory variables. Consequently the explanatory variables may be subject to experimental control, i.e. the experimental units may be selected on the basis of these variables. In the following the conditional expectations $E(Y|x)$, $E(Y_*|x)$ and $E(Y|x,Y_*)$ will be denoted as $E(Y)$, $E(Y_*)$ and $E(Y|Y_*)$.

Error-terms δ and ε are defined by

$$\delta = Y_* - E(Y_*), \quad \varepsilon = Y - E(Y|Y_*) \qquad \ldots(3)$$

With index j=1,2,...,N for the experimental units and j=1,2,...,n corresponding to the sub-sample where Y is observed, we have from first principles:

$$E(\delta_j) = 0, \quad var(\delta_j) = \sigma_0^2 \qquad \ldots(4)$$
$$E(\varepsilon_j) = 0, \quad var(\varepsilon_j) = \sigma_1^2$$
$$\varepsilon_j, \delta_{j'}, \; j \neq j' \text{ independent}, \quad cov(\varepsilon_j, \delta_j) = 0,$$
$$\varepsilon_j, \varepsilon_{j'}, \; j \neq j' \text{ independent}, \; \delta_j, \delta_{j'}, \; j \neq j' \text{ independent}$$

Furthermore

$$A = \alpha + \gamma a, \quad B = \beta + \gamma b \qquad \ldots(5)$$
$$\sigma^2 = \sigma_1^2 + \gamma^2 \sigma_0^2 \qquad \ldots(6)$$

$$\rho = \gamma\sigma_0/\sigma, \qquad \ldots(7)$$

where ρ is the partial correlation between Y and Y_* given x, i.e. the correlation between Y and Y_* conditional upon x.

We will assume that the errors follow a normal distribution

$$(\underline{\varepsilon}', \underline{\delta}')' \sim N(\underline{0}, \operatorname{diag}(\sigma_1^2 I_n, \sigma_0^2 I_N)) \qquad \ldots(8)$$

where $\underline{\varepsilon}' = (\varepsilon_1 \ldots \varepsilon_n)$ and $\underline{\delta}' = (\delta_1 \ldots \delta_N)$.

Assumption (8) simplifies the discussion but may be replaced by less restrictive regularity conditions without loss of the essential asymptotic properties of the estimators presented in 2.1.
Hence, the estimation procedure may be expected to be fairly robust against departures from normality. Non-linear extensions of (1) and (2) are discussed in section 6.

2.1 The estimators

a, b, σ_0^2 and α, β, γ, σ_1^2 may be estimated from the regression of Y_* on x and of Y on x and Y_* respectively by the method of least squares.
Hence, $\tilde{a}$ and $\tilde{b}$ are minimizing

$$S_0(a, b) = \Sigma(Y_* - a - bx)^2$$

and $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\gamma}$ are minimizing

$$S_1(\alpha, \beta, \gamma) = \Sigma(Y - \alpha - \beta x - \gamma Y_*)^2.$$

The variance estimators are

$$\tilde{\sigma}_0^2 = S_0(\tilde{a}, \tilde{b})/N-2) \quad \text{and} \quad \tilde{\sigma}_1^2 = S_1(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})/(n-3).$$

The following estimators for A, B and σ^2 are proposed:

$$\tilde{A} = \tilde{\alpha} + \tilde{\gamma}\tilde{a}, \quad \tilde{B} = \tilde{\beta} + \tilde{\gamma}\tilde{b}$$
$$\tilde{\sigma}^2 = \tilde{\sigma}_1^2 + \tilde{\gamma}^2\tilde{\sigma}_0^2. \qquad \ldots(9)$$

2.2 Properties of the estimators

Under the normality assumption (8) the logarithm of the (conditional) likelihood is

$$L \propto -N \log \sigma_0 - n \log \sigma_1 - {}^1/_2 S_0/\sigma_0^2 - {}^1/_2 S_1/\sigma_1^2 + \log \omega,$$

where ω follows from the procedure adopted for selecting the subsample of size n out of the total of N experimental units. When for instance the subsample is

taken at random: $\omega = \binom{N}{n}^{-1}$. We will assume that ω does not depend on the unknown location and scale parameters.

It follows that L is maximised for $\tilde{a}$, $\tilde{b}$, $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$, $\frac{N-2}{N}\tilde{\sigma}_0^2$ and $\frac{n-3}{n}\tilde{\sigma}_1^2$ as values for the parameters a, b, α, β, γ, σ_0^2 and σ_1^2.

Consequently $\tilde{A}$ and $\tilde{B}$ from (9) are maximum likelihood estimators. The maximum likelihood estimator for σ^2 is

$$\frac{n-3}{n}\tilde{\sigma}_1^2 + \frac{N-2}{N}\tilde{\gamma}^2\tilde{\sigma}_0^2.$$

The estimators $\tilde{a}$, $\tilde{b}$, $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$, $\tilde{A}$, $\tilde{B}$, $\tilde{\sigma}_0^2$ and $\tilde{\sigma}_1^2$ are unbiased. Both the maximum likelihood estimator for σ^2 and the estimator proposed in (9) are biased with a bias of order 1/n.
A bias corrected estimator for σ^2 is $\tilde{\sigma}^2 - \tilde{\sigma}_0^2\,\tilde{Var}(\tilde{\gamma})$, where $\tilde{Var}(\tilde{\gamma})$ is the familiar estimator for the conditional variance of $\tilde{\gamma}$ in the regression of Y on x and Y_*. However, this estimator may be negative. Replacing negative values by zero, again introduces a bias of order 1/n.
$(\tilde{a}, \tilde{b})'$ follows a normal distribution and $\tilde{\sigma}_0^2$ and $\tilde{\sigma}_1^2$ are independently distributed as $\sigma_0^2\,\chi^2_{N-2}/(N-2)$ and $\sigma_1^2\,\chi^2_{n-3}/(n-3)$ respectively, where χ^2_ν denotes a chi-square distribution with ν degrees of freedom.

For the other properties discussed we will resort to large sample theory, some details are given in appendix A.
Subject to regularity conditions, the distribution of $(\tilde{A}, \tilde{B})$ may be approximated by a normal distribution with mean (A, B) and variance-covariance matrix V, where

$$V = \begin{pmatrix} V_A & C_{AB} \\ C_{AB} & V_B \end{pmatrix}$$

and

$$\begin{aligned} V_A &= \tilde{Var}(\tilde{\alpha}) + \tilde{\gamma}^2\tilde{Var}(\tilde{a}) + \tilde{a}^2\tilde{Var}(\tilde{\gamma}) + 2\tilde{a}\,\tilde{cov}(\tilde{\alpha}, \tilde{\gamma}) \\ V_B &= \tilde{Var}(\tilde{\beta}) + \tilde{\gamma}^2\tilde{Var}(\tilde{b}) + \tilde{b}^2\tilde{Var}(\tilde{\gamma}) + 2\tilde{b}\,\tilde{cov}(\tilde{\beta}, \tilde{\gamma}) \qquad ...(10) \\ C_{AB} &= \tilde{cov}(\tilde{\alpha}, \tilde{\beta}) + \tilde{\gamma}^2\,\tilde{cov}(\tilde{a}, \tilde{b}) + \tilde{a}\tilde{b}\,\tilde{Var}(\tilde{\gamma}) + \tilde{a}\,\tilde{cov}(\tilde{\gamma}, \tilde{\beta}) + \tilde{b}\,\tilde{cov}(\tilde{\gamma}, \tilde{\alpha}) \end{aligned}$$

$\tilde{Var}(\tilde{\alpha})$, $\tilde{Var}(\tilde{a})$,... follow from the regressions of Y on x and Y_* and of Y_* on x, respectively.
It follows in particular that $\tilde{A}$ and $\tilde{B}$ are consistent ($n \to \infty$). From the distributional properties of $\tilde{\sigma}_0^2$, $\tilde{\sigma}_1^2$ and $\tilde{\gamma}$ it follows that $\tilde{\sigma}^2$ is a consistent estimator for $\sigma^2 (n \to \infty)$. Asymptotic normality may be derived for $\tilde{\sigma}^2$, but in

view of the skewness of the distribution the following approximation by a multiple of a chi-square distribution may be more appropriate:

$$\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{\tilde{\nu}}/\tilde{\nu} \qquad \ldots(11)$$

where $\tilde{\nu} = \tilde{\sigma}^4/\{\tilde{\sigma}_1^4/(n-3) + \tilde{Y}^4\tilde{\sigma}_0^4/(N-2) + 2\tilde{\sigma}_0^4\tilde{Y}^2\ \tilde{\mathrm{Var}}(\tilde{Y})\}$.

The denominator of the approximate degrees of freedom in (11) is half of the large sample variance of $\tilde{\sigma}^2$. $(1-\alpha)$ confidence intervals will approximately be a factor $\tilde{\nu}^2/(\tilde{\nu} - \frac{1}{2} - \frac{1}{2}u)^2$ larger than intervals derived under the normal approximation, where u is the $1-\alpha/2$ percentile point of the standard normal distribution.

3. EFFICIENCY RELATIVE TO DIRECT REGRESSION

In the following sections the estimation procedure introduced in section 2 will be referred to as 'double-regression'. In this section 'double-regression' based on a sample and sub-sample of sizes N and n respectively will be compared with direct regression of Y on x based on a sample of size m. The comparison will be based on large-sample results, some details are given in appendix B.

Under the following limiting conditions

$$\lim_{n\to\infty} n/N = f, \quad \lim_{n\to\infty} n/m = h$$

$$f \text{ and } h \text{ constant}, \quad 0 < f,h \leq 1,$$

the asymptotic relative efficiency for a parameter θ, denoted by ARE (θ) will be defined as the ratio of the asymptotic variances of the unbiased estimators of θ under direct regression and double-regression.
For k=1, both for the intercept A and slope B:

$$\mathrm{ARE}(A) = \mathrm{ARE}(B) = h/(1 - (1-f)\rho^2), \qquad \ldots(12)$$

where the partial correlation ρ is given in (7).

For the pair (A, B) we define ARE(A,B) as the square root of the ratio of the determinants of the asymptotic variance-covariance matrices under direct- and double-regression. ARE(A,B) is also given by (12).

For $k \geq 2$ we take the (k+1)th root of the ratio of the determinants as a definition for ARE$(A, B_1, \ldots, B_k)$. Again we find expression (12).

Obviously for h=1, i.e. n=m, the double-regression method will be the most efficient for large samples. For small samples however, this does not necessarily hold. For the particular case that β=b=0, a small-sample theory

based on maximum likelihood under normality has been developed by Conniffe and Moran (1972). The efficiency for the remaining parameter A is

$$h/\{1 - (1-f)\rho^2 + (1-f)(1-\rho^2)/(n-3)\}, \qquad \ldots(13)$$

where $h = n/m$.

For $h=1$ the direct method will be better when $\rho^2 < 1/(n-2)$. So for moderately sized n and small $|\rho|$ the direct method may be more efficient, because the observations on the concomitant variable may introduce more 'noise' than 'information'.
Simulation results in section 5 indicate that in practice this will be uncommon.

Expression (12) may be used to see how much there is too be gained in increase of precision at equal cost or in reduction of expense at equal precision, by replacing a direct regression experiment by a double-regression experiment.

4. APPLICATION TO THE PREDICTION OF LEAN MEAT PERCENTAGES

To establish a regression formula to predict the percentage of lean meat of pig carcasses in the Netherlands, an experiment was conducted where 200 carcasses were dissected by the IVO-standard method (Y_*). From these 200 carcasses 20 were chosen to proceed dissection from the joints to the EC-reference method (Y). For all carcasses the backfat and muscle thickness at the third to fourth from last rib position, 6 cm from the dorsal midline, were obtained with the Hennesy Grading Probe (x_1 and x_2). Some of the results are:

$(\tilde{a}, \tilde{b}_1, \tilde{b}_2) = (65.64, -0.6762, 0.0903)$, $\tilde{\sigma}_0^2 = 3.20$ with 197 degrees of freedom

$(\tilde{\alpha}, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\gamma}) = (-12.3, -0.0564, 0.0711, 1.079)$, $\tilde{\sigma}_1^2 = 0.694$ with 16 degrees of freedom

$\tilde{\rho} = 0.92$.

From (9):

$(\tilde{A}, \tilde{B}_1, \tilde{B}_2) = (58.53, -0.79, 0.17)$

$\tilde{\sigma}^2 = 4.42$ with $\tilde{\nu} = 26.35$ approximate degrees of freedom.

The bias corrected estimate for σ^2 is 4.08.

The coefficients $\tilde{\beta}_1$ and $\tilde{\beta}_2$ with standard errors 0.09 and 0.04 respectively are not significantly different from zero.

Replacing the regression of Y on Y_*, x_1 and x_2 by the regression of Y on Y_* only we have:

$(\tilde{A}, \tilde{B}_1, \tilde{B}_2) = (61.33, -0.76, 0.10)$
$\tilde{\sigma}^2 = 4.82$ with 127.24 approximate degrees of freedom.

The bias corrected estimate for σ^2 is 4.81.
Dropping x_1 and x_2 from the regression of Y on Y_*, x_1 and x_2 has some effect on the coefficient $\tilde{B}_2$ for muscle thickness. We return to this point later on in section 6.

From (12) we see that for a direct regression experiment with about equal precision, m=83 carcasses should be fully dissected according to the EC-reference method. This result follows by putting (12) equal to one with $f = 20/200 = 0.1$, $h = n/m = 20/m$ and $\rho = 0.92$.

5. SOME MONTE CARLO RESULTS

To see how the asymptotic results from sections 2 and 3 stand up for small samples a simulation study was performed. Some of the results will be discussed briefly in this section.

In the simulation k=1, N=200, n=20. For the explanatory variable x the values for backfat-thickness from the lean meat data were used. For each of four configurations of the parameters (see table 1) 1000 experiments were simulated with Genstat (Alvey et al., 1982).

Configuration I resembles the practical problem from section 4 except that muscle thickness was not included as an explanatory variable.
Configuration II has a large value for β, to study the effect of dropping x from the regression of Y on x and Y_*.

III and IV represent configurations with a fairly low and an almost negligible value for the partial correlation ρ.

Table 1: Four configurations of parameter values.

configuration	a	b	σ_0	α	β	γ	σ_1	A	B	σ^2	ρ
I	65.64	-0.68	1.79	-12.3	0.06	1.08	0.83	58.6	-0.674	4.43	0.92
II	65.64	-0.68	1.79	-12.3	0.60	1.08	0.83	58.6	-0.134	4.43	0.92
III	65.64	-0.68	0.57	-12.3	0.06	1.08	1.53	58.6	-0.674	2.72	0.37
IV	65.64	-0.68	0.57	-12.3	0.06	0.108	1.53	-5.21	-0.001	2.34	0.04

To see if the approximations for variances and covariances derived from asymptotic arguments give a fair impression of the accuracy of the estimators $\tilde{A}$ and $\tilde{B}$ for small samples, the actual coverage probabilities were determined

for the 95%-confidence intervals $\tilde{A} \pm 1.96 \sqrt{V_A}$, $\tilde{B} \pm 1.96 \sqrt{V_B}$ and for the 95%-confidence ellipsoid $(\tilde{A}-A, \tilde{B}-B)V^{-1} (\tilde{A}-A, \tilde{B}-B)' < 5.99$. For each run the elements of the variance-covariance matrix V were determined from (10).

The actual levels attained varied (standard errors between brackets) from 93.4 (0.79)% to 96.5 (0.58)% for the intervals and from 92.1 (0.85)% to 95.0 (0.69)% for the ellipsoid. Although results may be liberal, for all practical purposes the actual levels attained are satisfactory and the large-sample approximations give a reasonable impression of the accuracy of the estimators.

Defining the small sample efficiency of direct regression on the subsample only relative to double-regression as the ratio of sample variance and sample mean square error, results for A and B were of comparable size. Average values for configurations I, II, III and IV were 2.7, 3.6, 1.0 and 0.98 respectively. So even for $\rho = 0.04$ the double-regression method seems to be only slightly less efficient. For larger values of the partial correlation double-regression may be considerably more efficient.

Dropping x from the regression of Y on x and Y_* in configuration II, where β is substantially different from zero, results in poor coverage probabilities (below 75%) and a considerable bias in the estimators for A and B.

For configuration I from 2000 simulations coverage probabilities were determined for the large sample confidence interval for σ^2 under the normal and chi-square approximation respectively. The results were 92.8 (0.6)% and 94.0 (0.5)%, the chi-square approximation performing slightly better. Dropping x from the regression of Y on x and Y_* in configuration II results in a bias of -28.3% of the true value for the estimator of σ^2!

6. DISCUSSION

In this paper a method is presented, refered to as 'double-regression', to estimate regression coefficients, using additional information contained in a concomitant variable. When observations on the variable of interest are considerably more expensive than observations on the concomitant variable, the method allows:

- to cut down on the cost of the experiment without loss of efficiency relative to direct regression, or
- to increase the efficiency at the same cost.

The double-regression estimates and the approximate variances and co-variances are easily obtained from the output of any regression-package such as those contained in GENSTAT, GLIM, SAS, SPSS or BMD.

The double-regression procedure may be generalised by allowing the explanatory variables to enter the conditional expectations in (1) and (2) non-linearly. A product term of an explanatory variable and the concomitant variable Y_* may

be added to the conditional expectation of Y in (2).
In general a routine for non-linear regression will then be needed. These are supplied by many statistical computer-packages.

It seems crucial to the simplicity and robustness with respect to distributional assumptions of the double regression method that Y_* enters the conditional expectation of Y in (2) linearly. Since in practice Y and Y_* will often be measuring the same phenomenon, their close relationship will usually allow for a linear approximation, possibly after a suitable transformation of Y_*.

In Cook et al. (1983) analytical expressions for the optimal sample and subsample sizes for the estimators of Conniffe and Moran (1972) are derived. Similar (asymptotic) results may be derived for double-regression. An easy alternative is to simulate experiments for various sample sizes and a priori values of the parameters, to evaluate cost and precision and compromise between the two.

The Monte Carlo study indicates that combining the regressions of Y_* on $x_1,\ldots,x_k$ and of Y on Y_* (dropping $x_1,\ldots,x_k$) may give very poor results since the estimators for the regression coefficients and the residual variance may be seriously biased. So, although the procedure is an attractive one when the relationship between Y and Y_* has been established in the past and only 'cheap' observations on Y_* and $x_1,\ldots,x_k$ have to be collected, one should be very careful in replacing the regression of Y on Y_* and x by Y on Y_* only.

ACKNOWLEDGEMENTS

Assistance of W. Buist with the MC-study and careful reading of an earlier draft of this paper by P. Goedhart are gratefully acknowledged.

REFERENCES

Alvey, N., Galwey, N., Lane, P. (1982). An introduction to Genstat. Academic Press (152 pages).

Conniffe, D., Moran, M.A. (1972). Double sampling with regression in comparative studies of carcass composition. Biometrics, 28: 1011-1023.

Cook, G.L., Jones, D.W., Kempster, A.J. (1983). A note on a simple criterion for choosing among sample joints for use in double sampling. Anim.Prod., 36: 493-495.

Engel, B. (1987). Increasing precision or reducing expense in regression experiments by using information from a concomitant variable. Research report LWA-87-1 (25 pages).

Feller, W. (1966). An introduction to probability theory and its applications. Vol. II, John Wiley & Sons (626 pages).

APPENDICES

For a detailed account see Engel (1987) (available from the authors).

A Large sample distribution of $\tilde{A}$ and $\tilde{B}$

Let $m_n = \frac{1}{n}\sum_{j=1}^{n} x_j$ and $v_n = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$ (for $k \geq 2$ a vector and a matrix respectively).
Assume that: $\lim_{n\to\infty} m_n = m$ and $\lim_{n\to\infty} v_n = v$ exist and are finite with $v > 0$ (positive definite for $k \geq 2$).

Let X_1 be the design matrix for the regression of Y on x and Y_*. Then $\frac{1}{n} X_1'X_1$ converges in probability to a positive definite matrix, say $\lim_{n\to\infty} \frac{1}{n} X_1'X_1 = \Omega_1$. Let X_0 be the design matrix for the regression of Y_* on x. From similar assumptions: $\lim_{N\to\infty} \frac{1}{N} X_0'X_0 = \Omega_0$.

Collect the random elements of $\frac{1}{n} X_1'X_1$, $\frac{1}{n} X_1'Y$ and $\frac{1}{N} X_0'Y_*$ in a vector $\underset{\sim}{Z}$. Take any linear combination $\underset{\sim}{\lambda}'\underset{\sim}{Z}$ with $\underset{\sim}{\lambda} \neq 0$.
Asymptotic normality ($n\to\infty$) of $\underset{\sim}{\lambda}'\underset{\sim}{Z}$ follows from additional regularity conditions. For instance when the explanatory variable x is restricted to a bounded set, the Lindberg-condition for asymptotic (univariate) normality of a sum of independent random variables is met (see Feller (1966)). It follows that $\underset{\sim}{Z}$ is asymptotically (multivariate) normally distributed.

$\underset{\sim}{\tilde{\theta}} = (\underset{\sim}{\tilde{\theta}}_0', \underset{\sim}{\tilde{\theta}}_1')'$, with $\underset{\sim}{\tilde{\theta}}_0 = (\tilde{a}, \tilde{b})'$ and $\underset{\sim}{\tilde{\theta}}_1 = (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})'$, may be related to $\underset{\sim}{Z}$ by a first order approximation. From a Taylor series expansion for $(\frac{1}{n} X_1'X_1)^{-1}$ it may be shown that

$$\sqrt{n}\,\Psi(\underset{\sim}{\tilde{\theta}}_1 - \underset{\sim}{\theta}_1) = H_1\,\sqrt{n}(\underset{\sim}{Z} - E(\underset{\sim}{Z})) + O_p(n^{-1/2}),$$

where $\Psi = E(\frac{1}{n} X_1'X_1)$ and H_1 is a matrix of constants.

Furthermore $\sqrt{n}(\frac{1}{N} X_0'X_0)(\underset{\sim}{\tilde{\theta}}_0 - \underset{\sim}{\theta}_0) = H_0\,\sqrt{n}(\underset{\sim}{Z} - E(\underset{\sim}{Z}))$, where H_0 is a matrix of constants. From $\lim_{n\to\infty} \Psi = \Omega_1$ and $\lim_{n\to\infty} \frac{1}{N} X_0'X_0 = \Omega_0$ it follows that $\sqrt{n}(\underset{\sim}{\tilde{\theta}} - \underset{\sim}{\theta})$ is asymptotically normally distributed ($n\to\infty$).
Since $\mathrm{Var}(\sqrt{n}(\underset{\sim}{\tilde{\theta}}_1 - \underset{\sim}{\theta}_1) \,|\, \underset{\sim}{Y}_*) = (\frac{1}{n} X_1'X_1)^{-1}\sigma_1^2 \to \Omega_1^{-1}\sigma_1^2$, $n\to\infty$ and $\tilde{\theta}_0$, $\tilde{\theta}_1$ are uncorrelated, the variance of the asymptotic distribution of $\sqrt{n}(\underset{\sim}{\tilde{\theta}} - \underset{\sim}{\theta})$ is

$$\mathrm{diag}\,(f\,\Omega_0^{-1}\sigma_0^2,\ \Omega_1^{-1}\sigma_1^2).$$

The asymptotic distribution of $(\tilde{A}, \tilde{B})'$ follows from:

$$\sqrt{n}\,(\tilde{A}-A,\ \tilde{B}-B)' = \begin{pmatrix} \gamma & 0 & 1 & 0 & a \\ 0 & \gamma & 0 & 1 & b \end{pmatrix} \sqrt{n}\,(\tilde{\underline{\theta}}-\underline{\theta}) + O_p(n^{-1/2})$$

Replacing α, β, γ, a, b and the elements of Ω_0 and Ω_1 by consistent estimators results in (10).

B Efficiency

The general result is derived for $k \geq 1$.

Let $E(\underline{Y} \mid \underline{Y}_{*(n)}) = M\underline{\mu} + \underline{Y}_{*(n)}\gamma$
$E(\underline{Y}_*) = X_0\,\underline{\theta}_0$
$E(\underline{Y}) = M\,\underline{\xi}$

where $\underline{\xi} = \underline{\mu} + \gamma\underline{\theta}_0$, $\underline{\mu}$ contains the intercept and coefficients for $x_1,\ldots,x_k$ and $\underline{Y}_{*(n)}$ denotes the vector of values of the concomitant variable for the subsample. So $X_1 = (M, \underline{Y}_{*(n)})$, $\underline{\theta}_1 = (\underline{\mu}', \gamma)'$ and $\underline{\xi} = (A, B_1,\ldots,B_k)'$.

Let $\lim_{n\to\infty} \frac{1}{n} M'M = \lim_{N\to\infty} \frac{1}{N} X_0'X_0 = \Omega_0$, $\lim_{n\to\infty} \frac{n}{N} = f$, $0 < f \leq 1$.

Then the asymptotic variance $(n\to\infty)$ of $\tilde{\underline{\xi}}$ is

$$(1 - (1-f)\rho^2)\Omega_0^{-1}\sigma^2.$$

For direct regression with m observations let the design matrix be X_3.
Let $\lim_{n\to\infty} \frac{n}{m} = h$, $0 < h \leq 1$, $\lim_{m\to\infty} \frac{1}{m} X_3'X_3 = \Omega_0$.
Then for direct regression the asymptotic variance is

$$h\,\Omega_0^{-1}\,\sigma^2$$

The product of the inverse of the first asymptotic variance and the second asymptotic variance is $hI/(1-(1-f)\rho^2)$. Expression (12) follows by taking determinants on both sides of the equation and taking the (k+1)-th root.

Ontvangen: 02-09-1988
Geaccepteerd: 05-10-1988